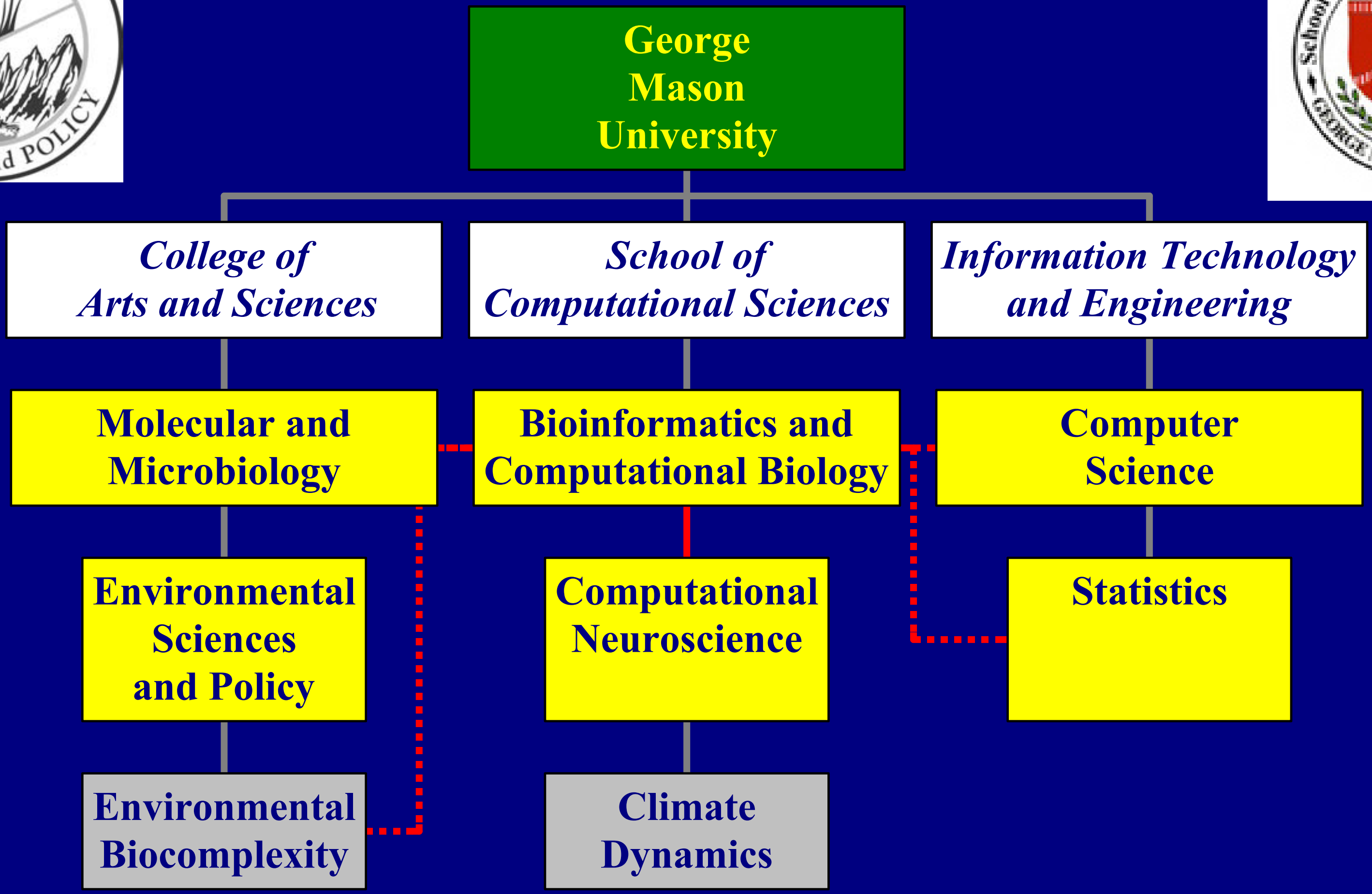


Linkage of Bioinformatics, Computer Sciences, and Experimental Sciences

Prince William Campus, Manassas, VA

Environmental Biocomplexity
Monitoring Microbial Communities

Patrick Gillevet

Objectives

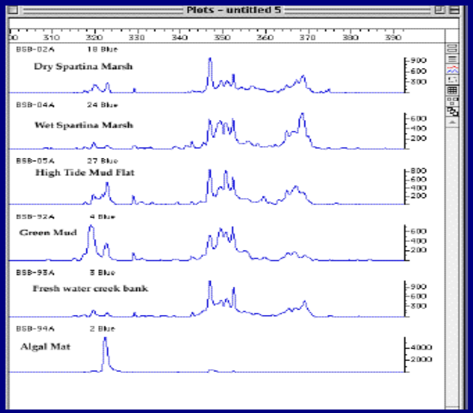
- Comprehensive continuous field monitoring architecture for bacterial, protists, and fungi
- Understand behavior of groups of organisms and their response to environmental change

Approach

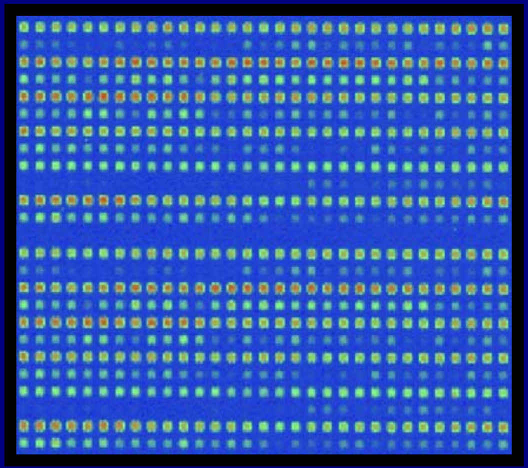
- Amplicon Length Heterogeneity (ALH)
  - Continuous monitoring of community dynamics
  - SSU rRNA, ITS, functional genes
- Sequencing SSU rRNA clone libraries
  - Initial characterization of community
- Microarray technology
  - Rapid characterization of samples of interest (verification)

Applications

- Risk assessment of genetically modified organisms
- Monitoring polymicrobial diseases -Crohn's disease
- Monitoring bioremediation processes
- Complete Background Characterization
- Monitoring Biowarfare agents & interferants



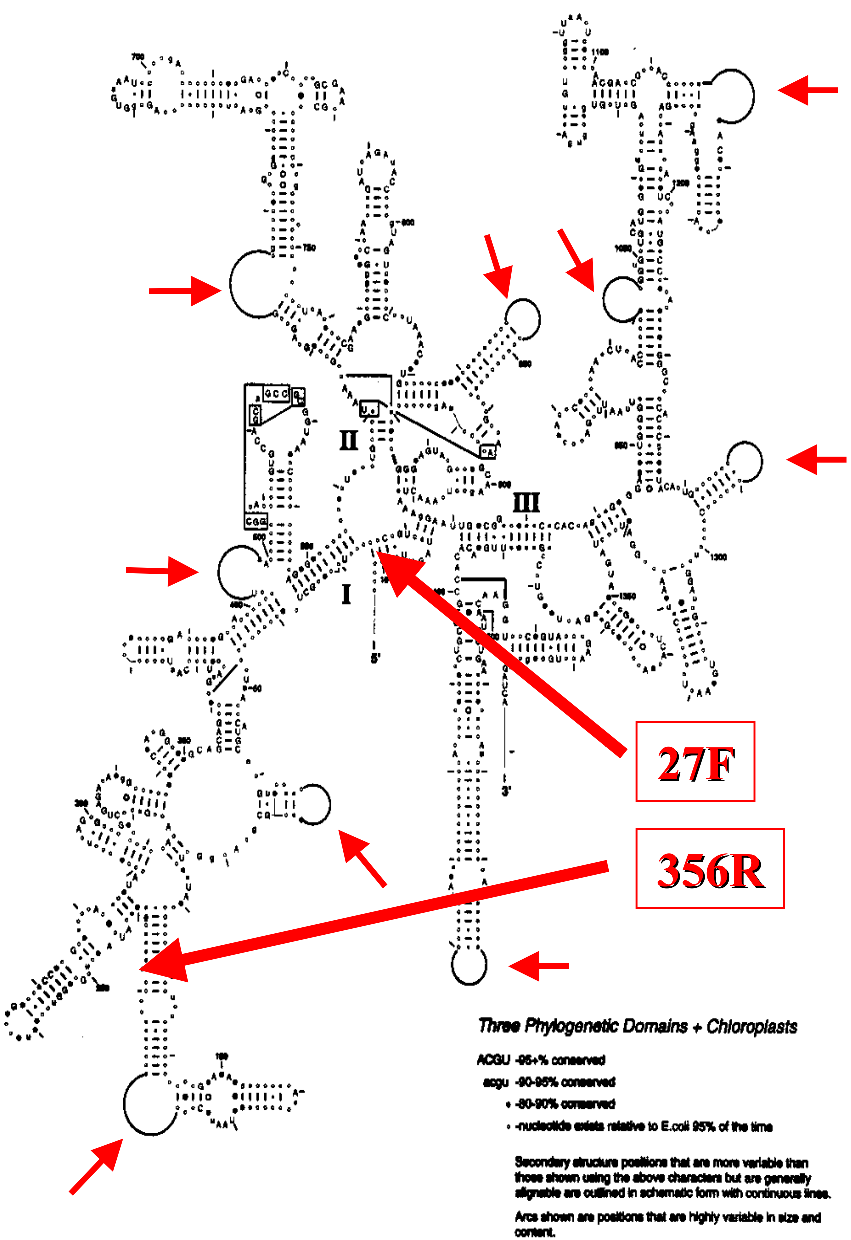
ALH Fingerprinting



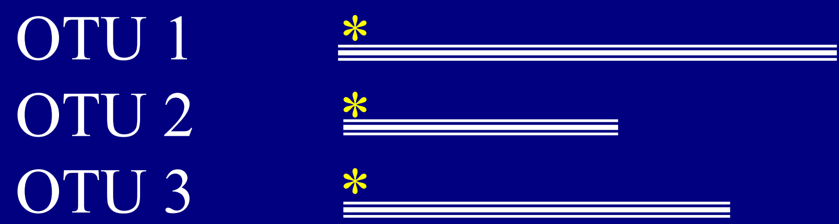
Oligo Microarrays

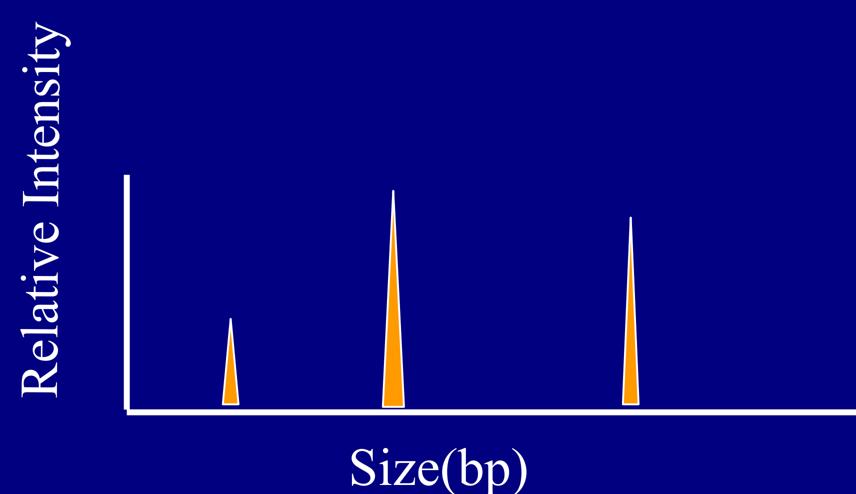
Sponsors: NSF, Sea Grant, CBNP

Phylogenetic conservation superimposed onto the *Escherichia coli* small subunit ribosomal RNA secondary structure



Amplicon Length Heterogeneity Fingerprinting (ALH)



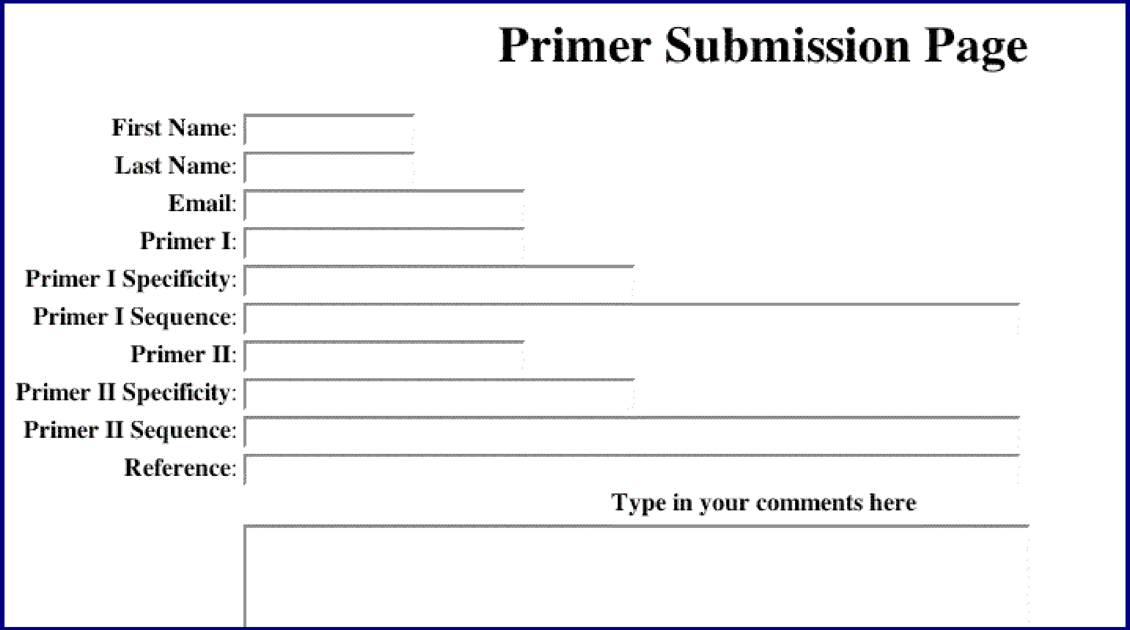


Peak area ~ Abundance

Amplcion Length Heretogeneity Analysis System (ALHAS)

P. Gillevet, J. Grefenstette, L. Kumar, A. Ahmed, A. Rehman

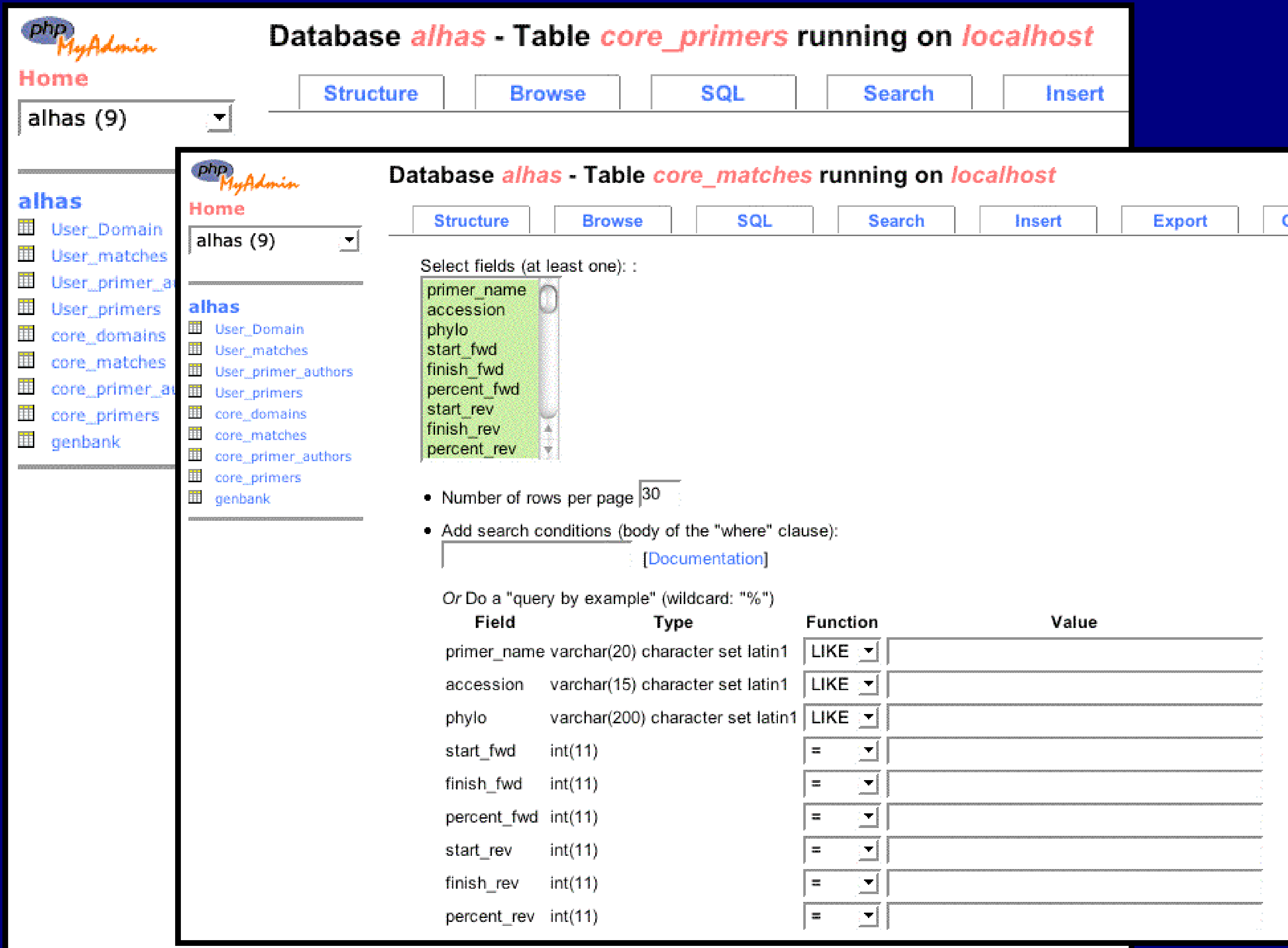
- Compile rRNA data from Genbank
- Calculate amplicon length (Domain) from primers
- Determine specificity of Domain
- Predict community composition from fingerprint



Primer Submission Page

First Name:
Last Name:
Email:
Primer I:
Primer I Specificity:
Primer I Sequence:
Primer II:
Primer II Specificity:
Primer II Sequence:
Reference:
Type in your comments here

http://binf.gmu.edu:16080/ALHAS/


Database alhas - Table core_primers running on localhost
Home
alhas (9)
Structure
Browse
SQL
Search
Insert
alhas
User_Domain
User_matches
User_primers
core_domains
core_matches
core_primers
genbank
Database alhas - Table core_matches running on localhost
Home
alhas (9)
Structure
Browse
SQL
Search
Insert
Export
alhas
User_Domain
User_matches
User_primer_authors
User_primers
core_domains
core_matches
core_primer_authors
core_primers
genbank
Select fields (at least one): :
primer_name
accession
phylo
start_fwd
finish_fwd
percent_fwd
start_rev
finish_rev
percent_rev
Number of rows per page 30
Add search conditions (body of the "where" clause):
[Documentation]
Or Do a "query by example" (wildcard: "%")
Field
Type
Function
Value
primer_name varchar(20) character set latin1 LIKE
accession varchar(15) character set latin1 LIKE
phylo varchar(200) character set latin1 LIKE
start_fwd int(11) =
finish_fwd int(11) =
percent_fwd int(11) =
start_rev int(11) =
finish_rev int(11) =
percent_rev int(11) =

Ecoinformatics Tools for Microbial Diversity Studies: Supervised Classification of Amplicon Length Heterogeneity (ALH) Profiles of 16S rRNA

Chengyong Yang, Yong Wang, DeEtta Mills, Kalai Mathee, Krish Jayachandran, Masoumeh Sikaroodi, Patrick Gillevet, Jim Entry, Giri Narasimhan
George Mason University and Florida International University

***Supervised classification* tools :**

(a) *Support Vector Machines* (SVM)
(b) K-*Nearest Neighbor Method* (KNN)

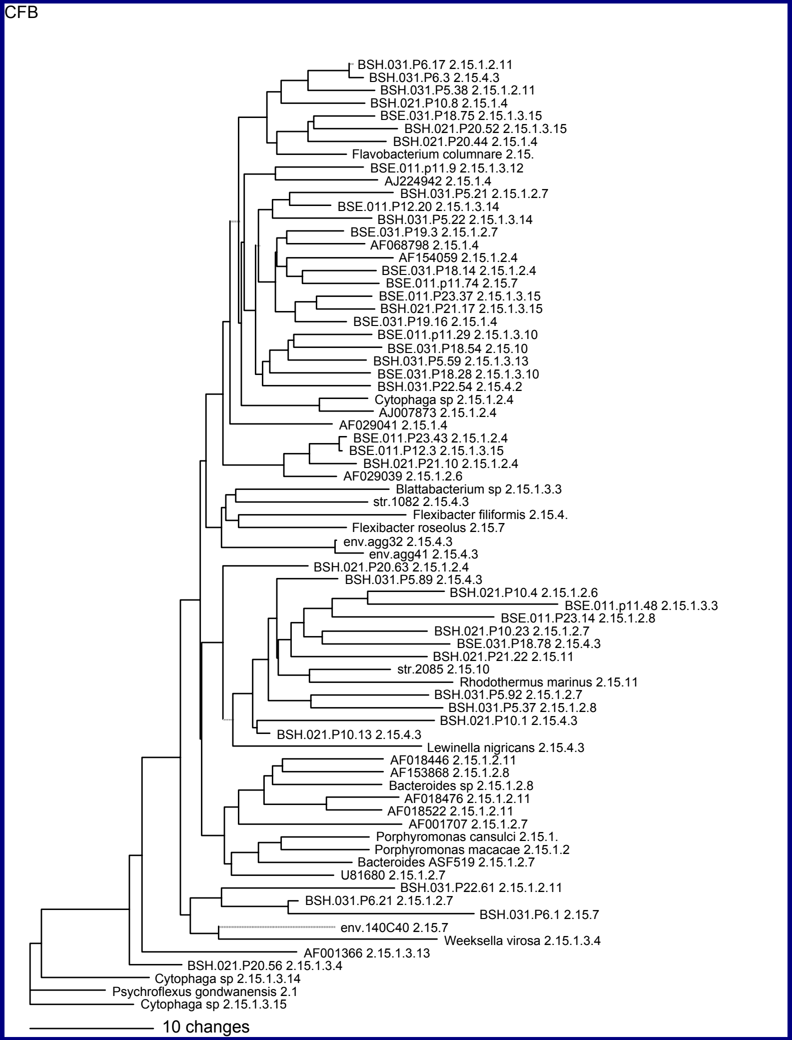
Analyzed 4 amplicons:	V1+V2	6-FAM-27F and 355R
	V1	6-FAM-P1F and P1R
	V3	HEX-338F and 518R
	V9	NED-1055F and EC1392R

SVM was able to classify Idaho soils but not Chesapeake Bay sediments
V1+V2 was most informative

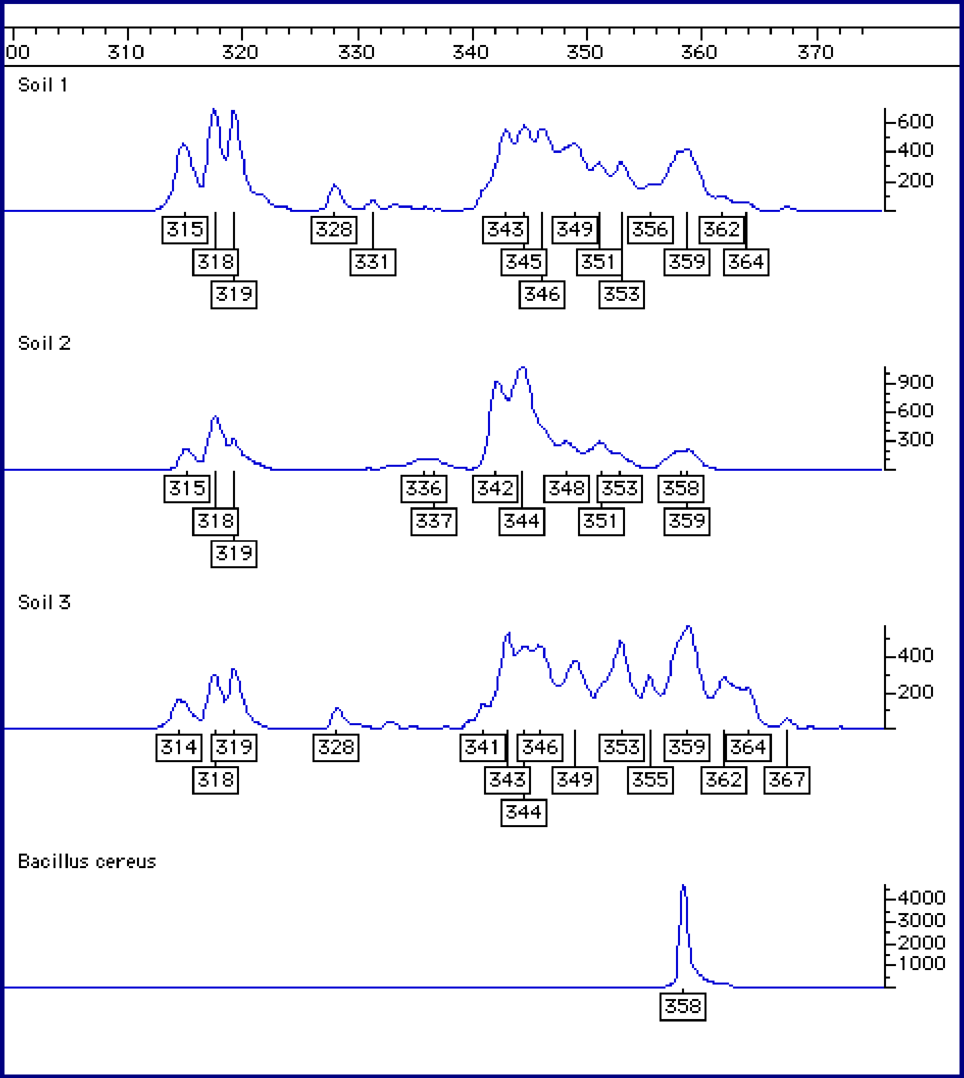
Automatically Megablast against RDP database
Sort by RDP Number

RDP number	Description	BSE_A	BSH_A	BSE_B	BSH_B	Total	Level 3/4 Total	LEVEL_1	LEVEL_2
1.2.2.1.4	ENVIRONMENTAL_CLONE_4B7_SUBGROUP				1	**1**	**1**	ARCHAEA	CRENARCHAEOTA
2.1	THERMOPHILIC_OXYGEN_REDUCERS				1	**1**		BACTERIA	THERMOPHILIC_OXYGEN_REDUCERS
2.3	CTM.PROTEOLYTICUS_GROUP				1	**1**	**2**	BACTERIA	CTM.PROTEOLYTICUS_GROUP
2.7.1.2	ENVIRONMENTAL_CLONE_T78_GROUP		1		1	**2**		BACTERIA	GREEN_NON-SULFUR_BACTERIA_AND_RELATIVES
2.7.2.1.3	MEI.RUBER_SUBGROUP	1				**1**	**3**	BACTERIA	GREEN_NON-SULFUR_BACTERIA_AND_RELATIVES
2.9.1.2	NSP.MOSCOVIENSIS_SUBGROUP				1	**1**		BACTERIA	LEPTOSPIRILLUM-NITROSPIRA
2.9.3.1	LPP.FERROOXIDANS_SUBGROUP				1	**1**		BACTERIA	LEPTOSPIRILLUM-NITROSPIRA
2.9.4	TDV.YELLOWSTONII_GROUP	1		1		**2**	**4**	BACTERIA	LEPTOSPIRILLUM-NITROSPIRA
2.10.1	ENVIRONMENTAL_CLONE_WCHB1-41_SUBGROUP			1		**1**		BACTERIA	PROSTHECOBACTER_GROUP
2.10.2	VER.SP_STR_VEGLC2_SUBGROUP			1		**1**	**2**	BACTERIA	PROSTHECOBACTER_GROUP
2.12	ENVIRONMENTAL_CLONE_OPB2_GROUP	1		1	1	**3**	**3**	BACTERIA	ENVIRONMENTAL_CLONE_OPB2_GROUP
2.13.1	ENVIRONMENTAL_CLONE_JAP604_GROUP		1			**1**		BACTERIA	NITROSPINA_SUBDIVISION
2.13.4	ENVIRONMENTAL_CLONE_OPB5_GROUP		1		1	**2**		BACTERIA	NITROSPINA_SUBDIVISION
2.13.5	ENVIRONMENTAL_CLONE_RB25_GROUP				1	**1**		BACTERIA	NITROSPINA_SUBDIVISION
2.13.6	ENVIRONMENTAL_CLONE_III1-8_GROUP			1		**1**		BACTERIA	NITROSPINA_SUBDIVISION
2.13.8	HP.FOETIDA_GROUP				1	**1**		BACTERIA	NITROSPINA_SUBDIVISION
2.13.12	ENVIRONMENTAL_CLONE_RB40_GROUP		1			**1**	**7**	BACTERIA	NITROSPINA_SUBDIVISION
2.14	FLS.SINUSARABICI_ASSEMBLAGE		1			**1**	**1**	BACTERIA	FLS.SINUSARABICI_ASSEMBLAGE
2.15.1.2.4	CY.FERMENTANS_SUBGROUP	1	2	1		**4**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.2.6	BAC.SPLANCHNICUS_SUBGROUP		1			**1**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.2.7	PPM.MACACAE_SUBGROUP		1	1	3	**5**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.2.8	BAC.FRAGILIS_SUBGROUP	1			1	**2**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.2.11	PRV.RUMINICOLA_SUBGROUP				3	**3**	**15**	BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.3.3	BLT.SP_SUBGROUP	1				**1**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.3.4	EMB.BREVIS_SUBGROUP		1			**1**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.3.10	F.FLEVENSE_SUBGROUP	1		1		**2**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.3.12	CAP.OCHRACEA_SUBGROUP	1				**1**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.3.13	CY.ULIGINOSA_SUBGROUP				1	**1**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.3.14	PSF.TORQUIS_SUBGROUP	1			1	**2**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.3.15	PSH.BURTONENSIS_SUBGROUP	2	2	1		**5**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.1.4	CYTOPHAGA_GROUP_II		2	1		**3**	**16**	BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.4.2	FLX.SANCTI_SUBGROUP				1	**1**		BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.4.3	LEW.NIGRICANS_SUBGROUP		2	1	2	**5**	**6**	BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.7	FLX.LITORALIS_GROUP	1			1	**2**	**2**	BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.8	TMN.LAPSUM_GROUP				1	**1**	**1**	BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES
2.15.10	STR.SBR2085			1		**1**	**1**	BACTERIA	FLEXIBACTER-CYTOPHAGA-BACTEROIDES

Automatic Report Generation



Survey of Backgrounds in Washington DC



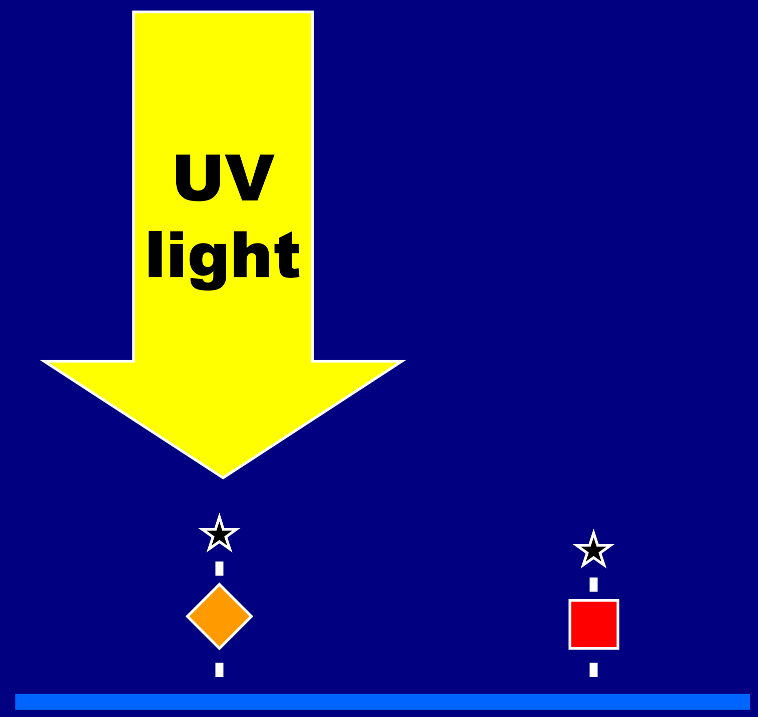
Identification of environment (soil 2) with least Biological Background

Oligo Synthesis Controlled by Light
2.5 orders of magnitude reduction in cost

Steve Smith

☆ Photo labile protecting group

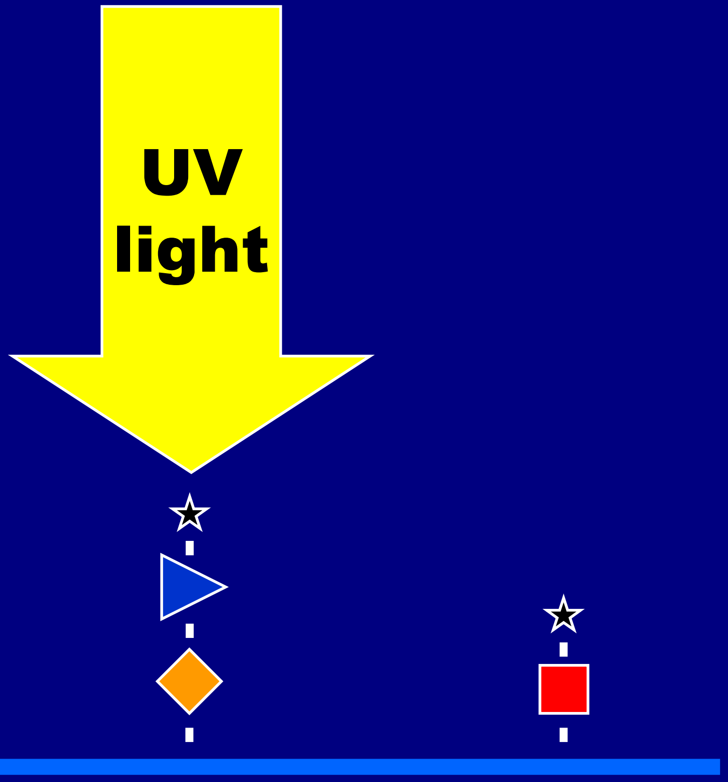
◆ DNA monomer



Oligo Synthesis Controlled by Light

☆ Photo labile protecting group

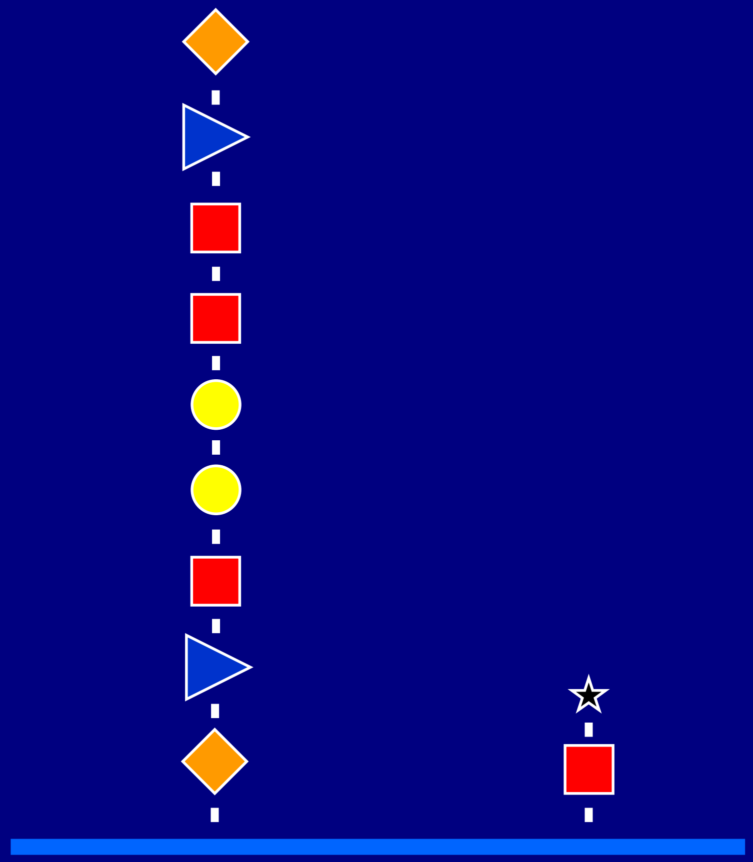
◆ DNA monomer



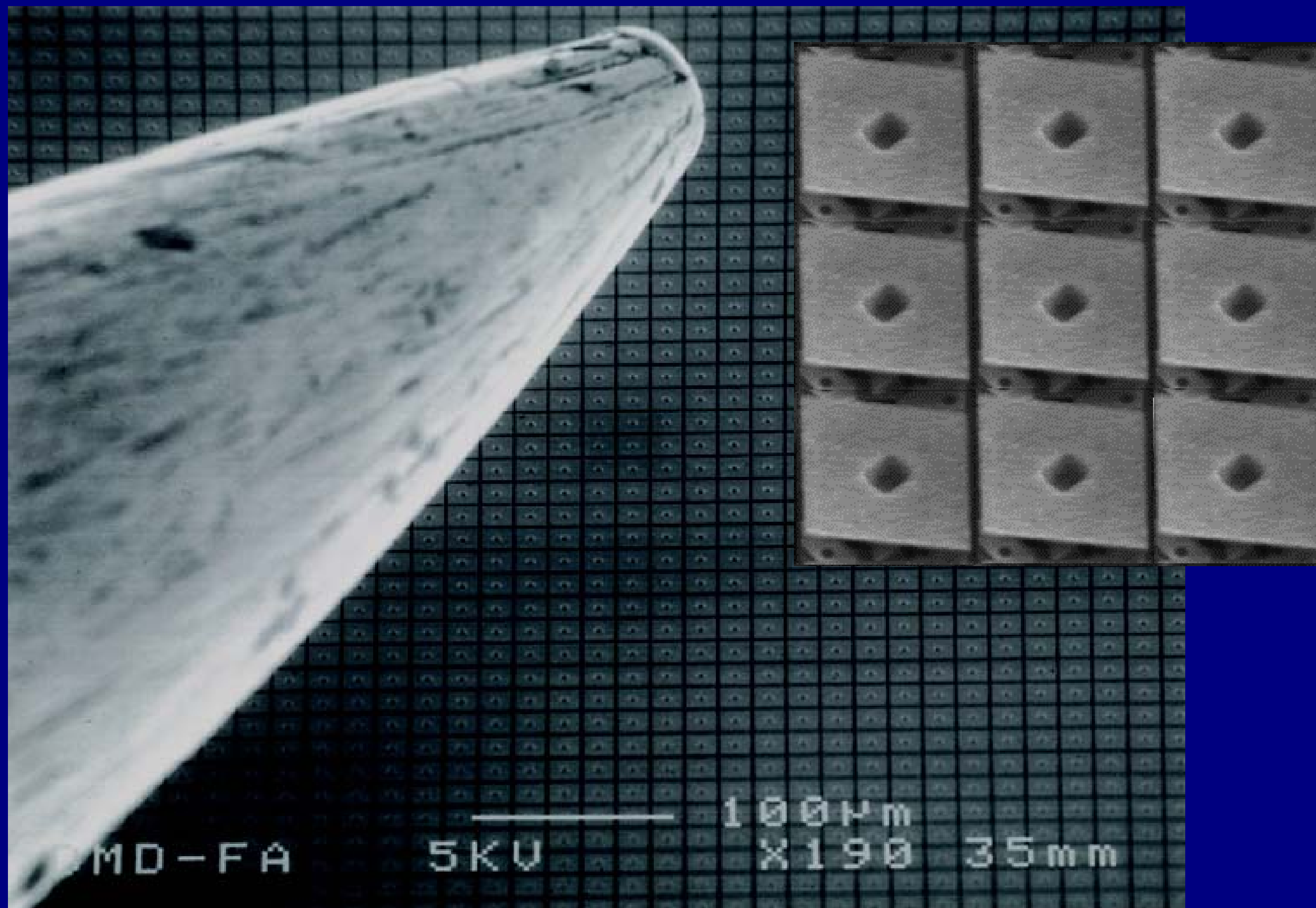
Oligo Synthesis Controlled by Light

☆ Photo labile protecting group

◆ DNA monomer



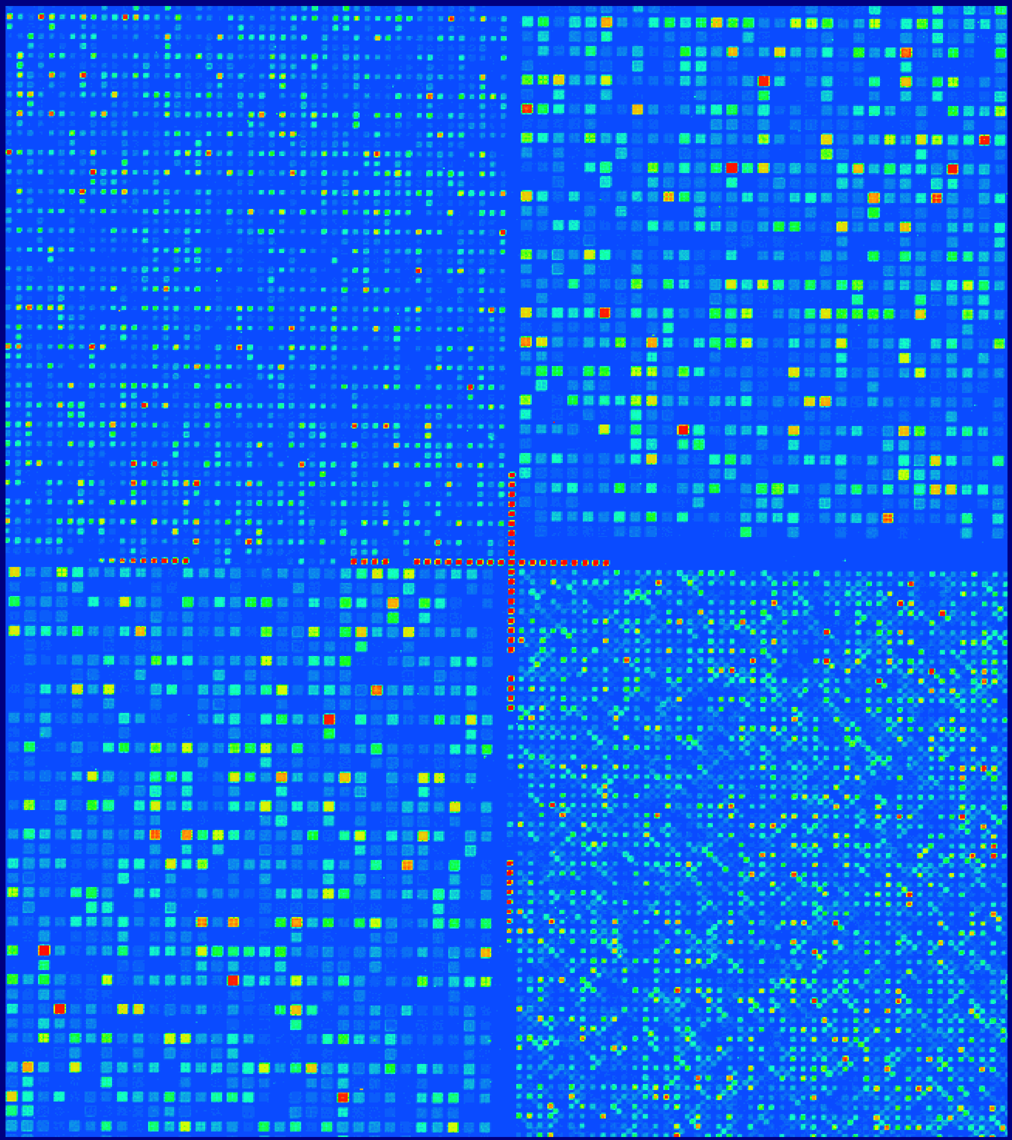
DMD: Digital Micromirror Device



195,000 (37mer - 70 mer)

390,000 (< 36mer)

Array Options



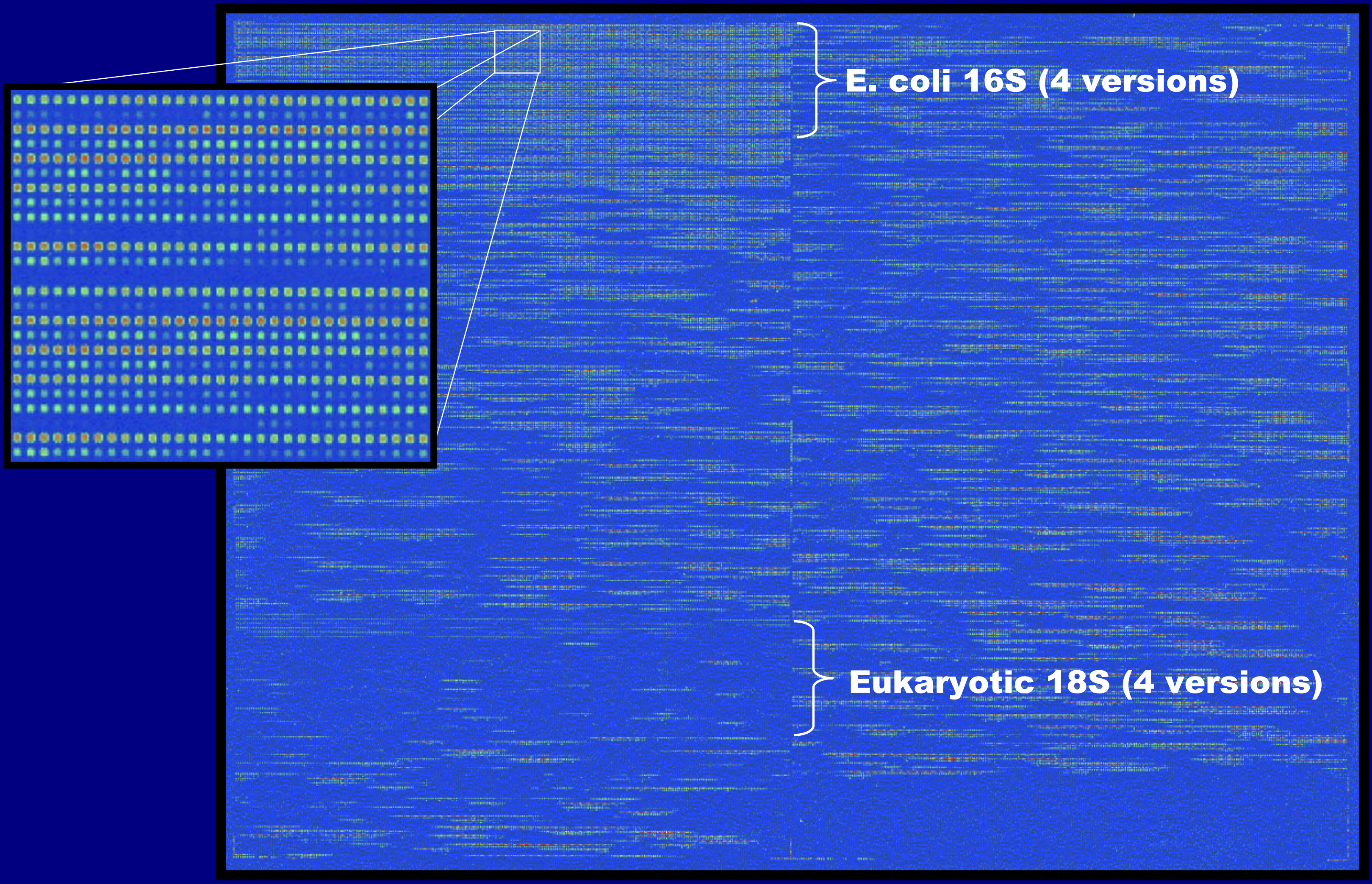
- **1 - 36mers**

390,000 features per array

- **37 - 70mers**

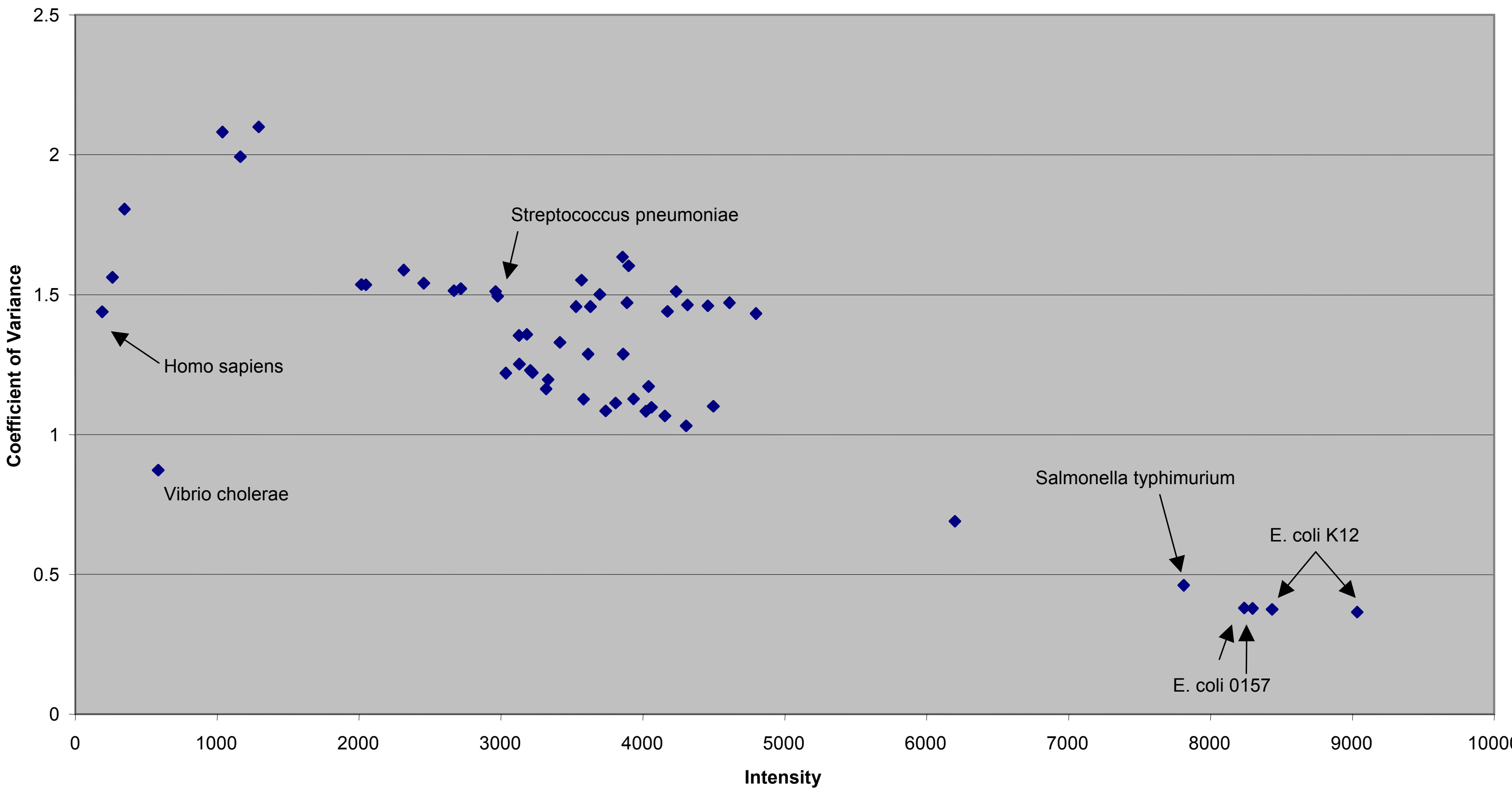
195,000 features per array

Tiling Ribosomal RNA of 53 Species

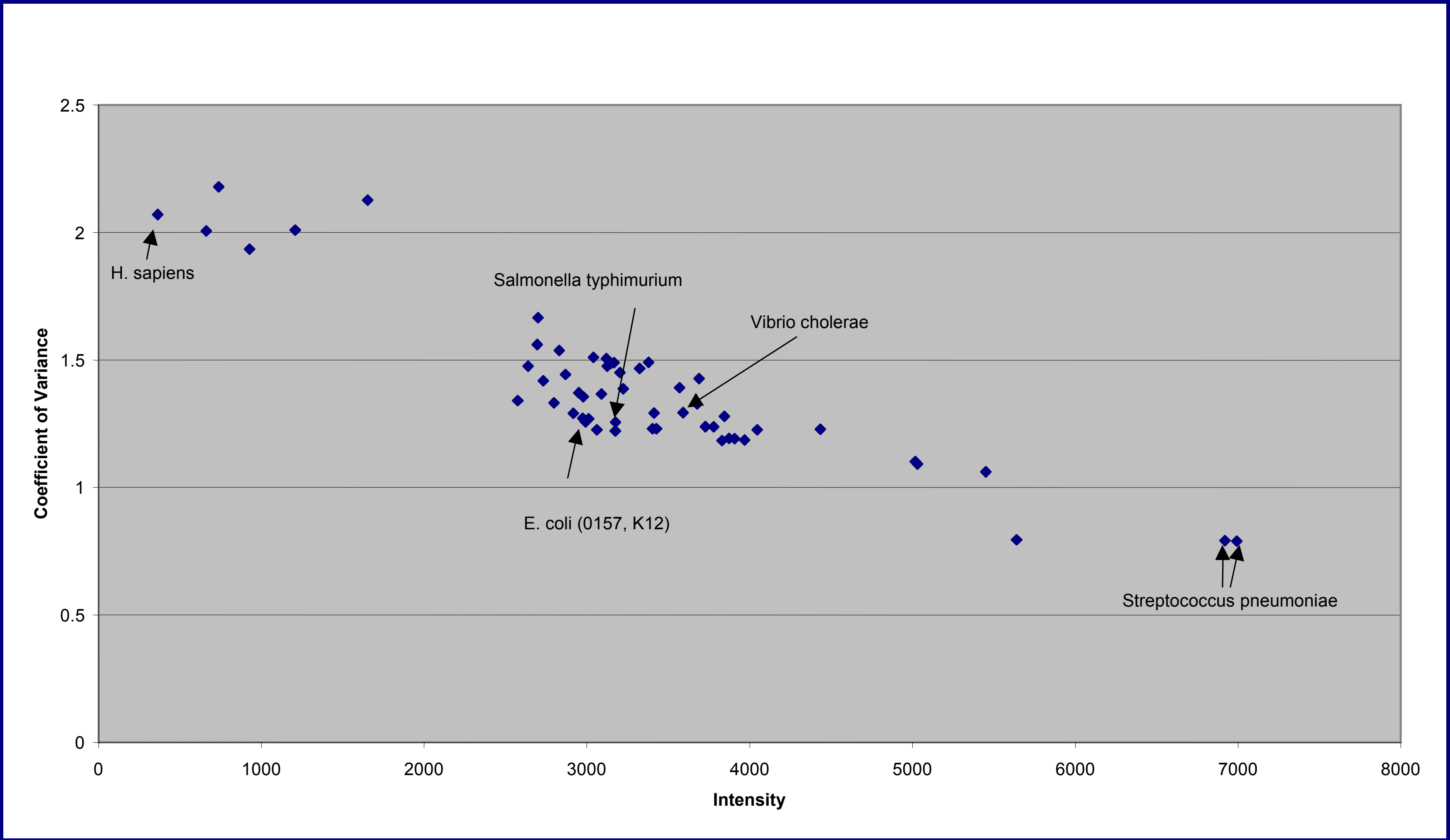


Hybridized to E. coli K12 RNA

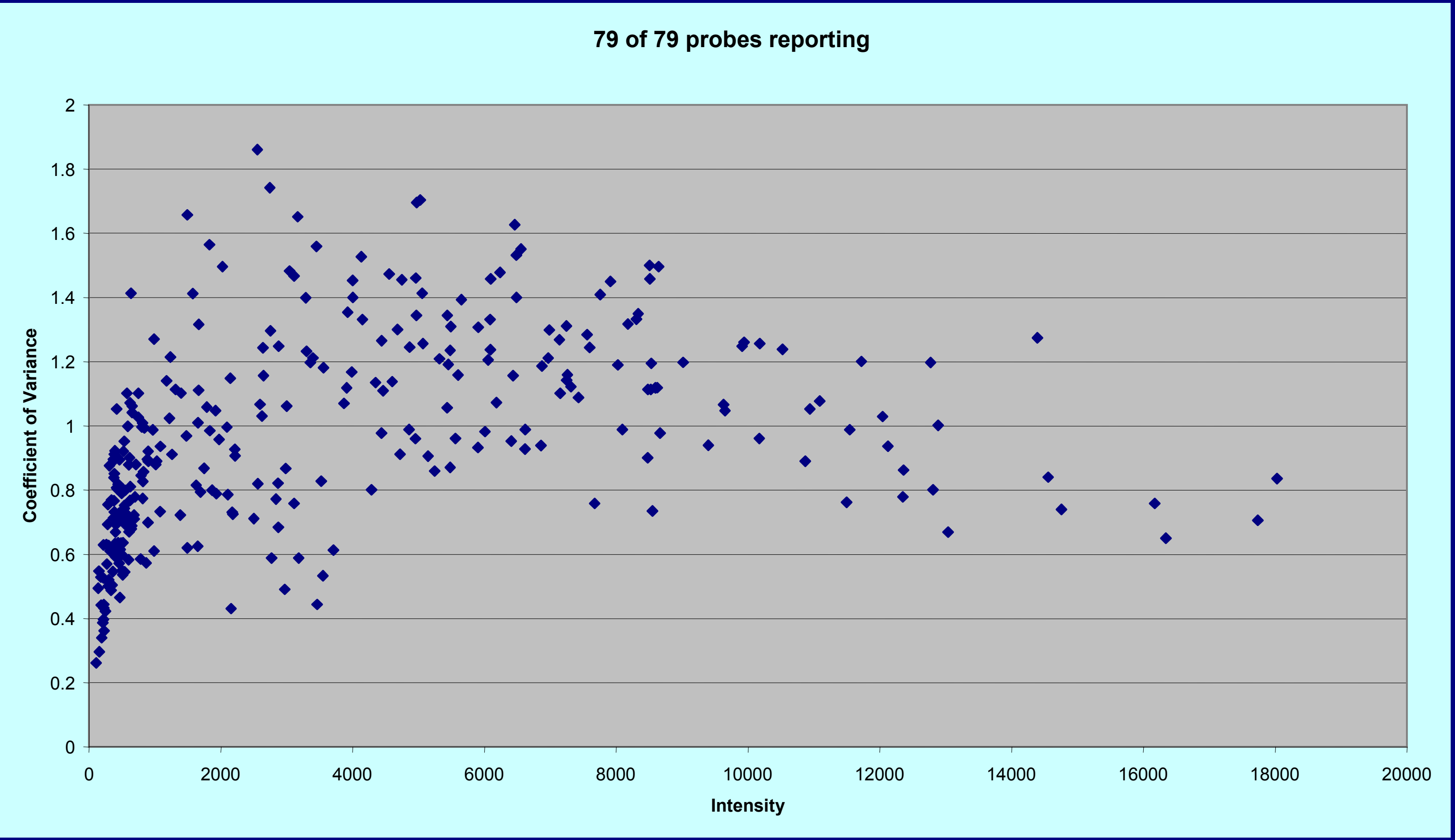
79 oligos x 55 Taxa Array Hybridized to E. coli K12 RNA



79 oligos to 55 Taxa Array
Hybridized to Oral Swab DNA



79 probes to 4816 Taxa Array
390,000 oligos
Hybridized to Oral Swab DNA



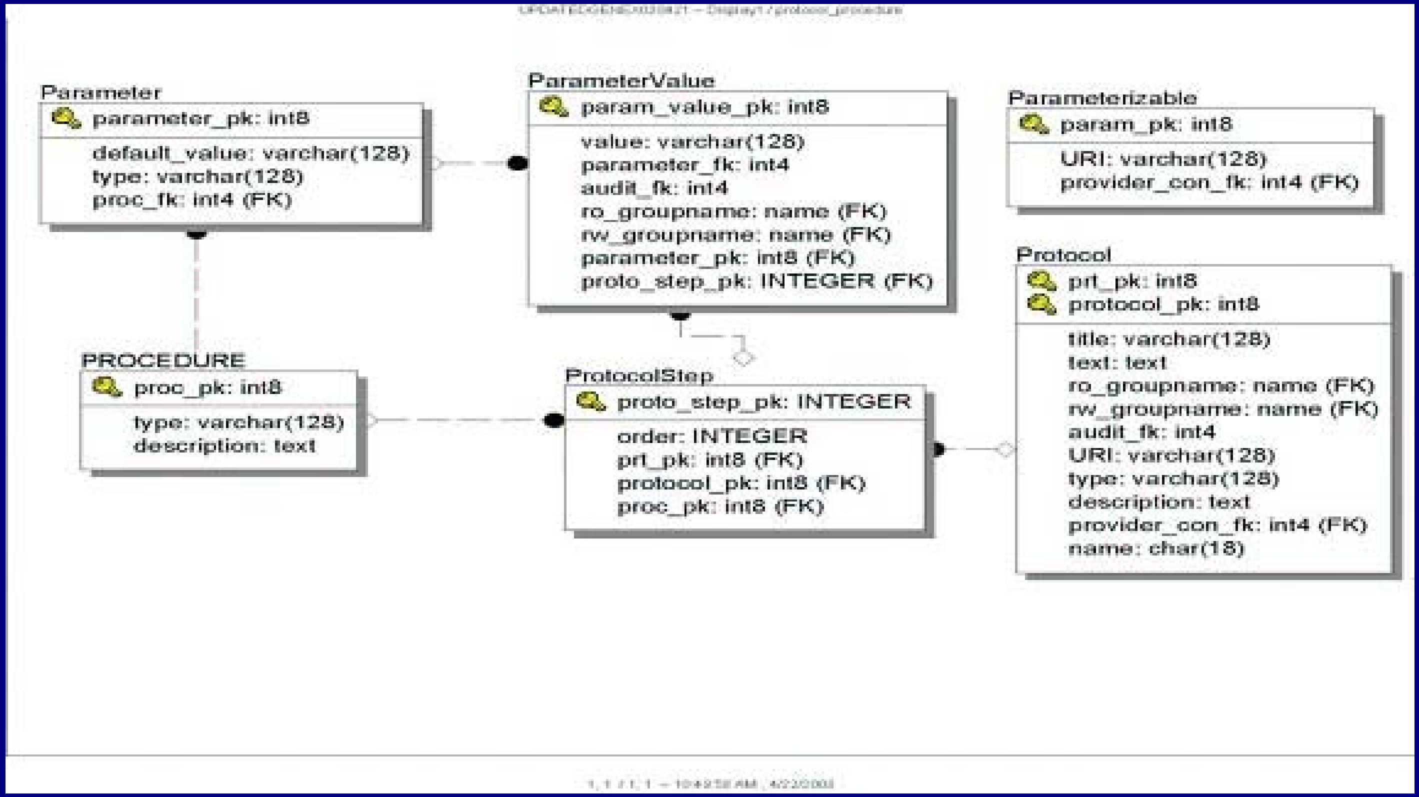
GMU: Possible use of supervised and unsupervised methods
Combine ALH and Microarray data to deconvolve community

Bioinformatics Solutions

Jennifer W. Weller

- Problem: Data from disparate sources are combined in one analysis. How do you track the source and the evidence- weight chosen in a particular analysis?
- Solutions
  – LIMS systems and e-notebooks (in concert with bar coding) track the processes of acquiring physical data.
  – Database handling of branched protocols (with method- or investigator- specific parameters) that can be automatically called.

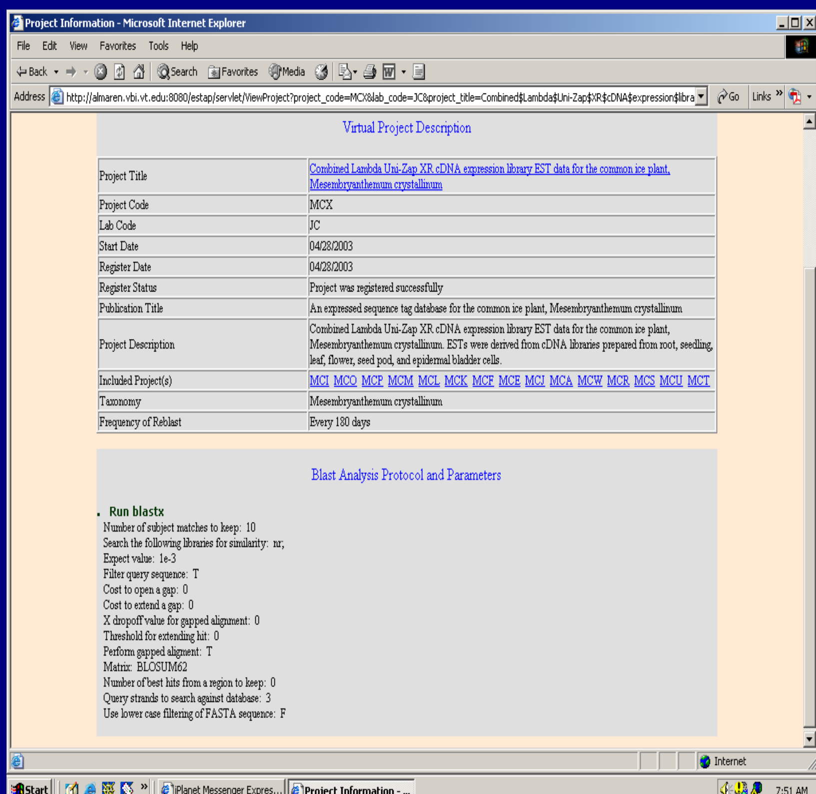
BioInformatics Database Solutions

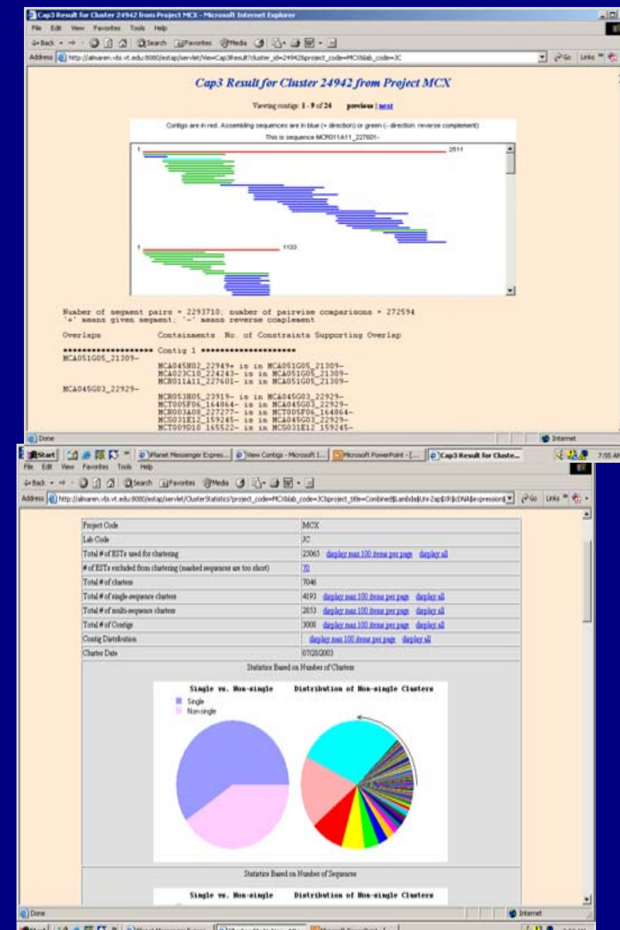


- The goal: track how data are transformed
- The basic components are reusable and reconfigurable
  - currently implemented using perl scripts for PostgreSQL

Current Implementations

- ESTAP : EST analysis pipeline



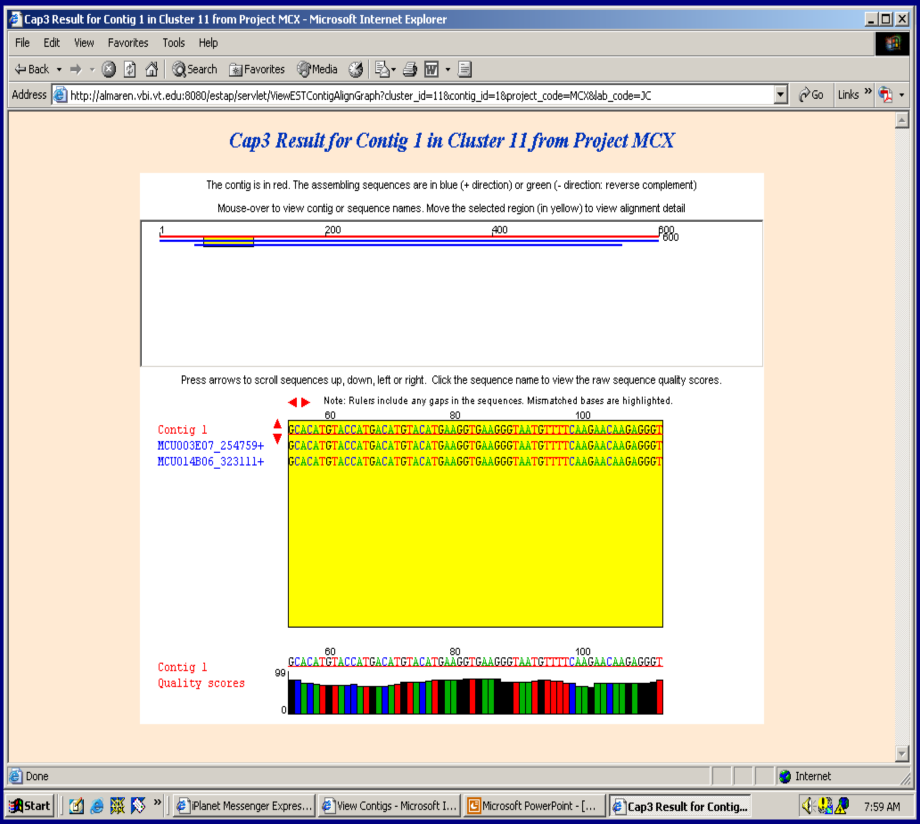


ESTAP continued

There are many ways to cluster ESTs.

SNP discovery comes from clustering – for some applications the quality score may be more rigorously controlled than for others.

The underlying data model allows parallel analyses to be run and stored, then called for a particular application.



GeneX Microarray Database

Jennifer Weller

Objectives

- Organize large microarray datasets for sharing, analysis and data mining
- Develop analysis methods that provide insight into biological processes based on microarray data

Approach

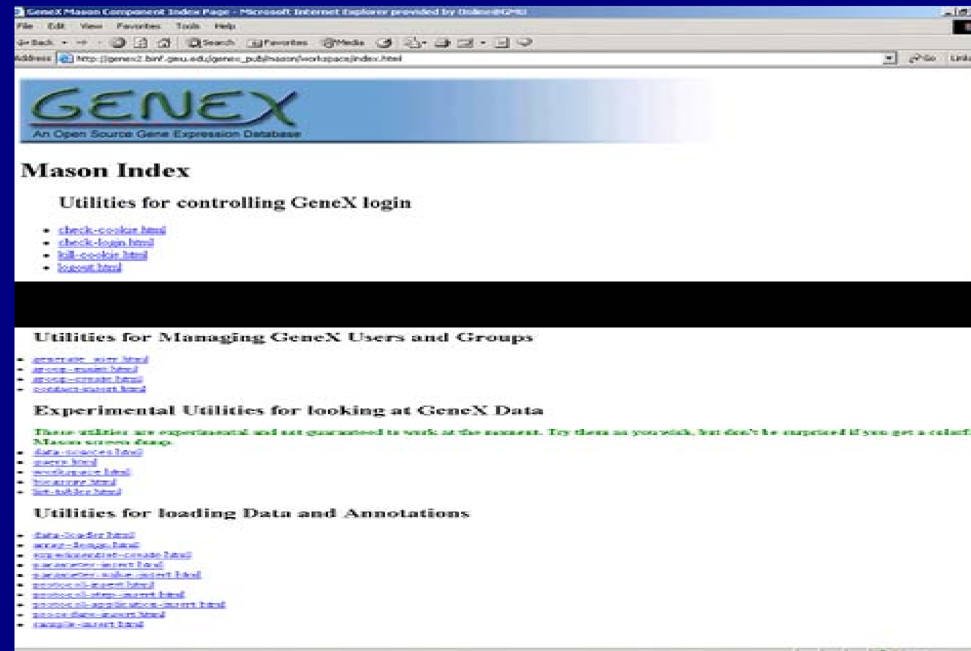
- Develop a reference implementation of the MAGE model (GeneX)
- Populate the GeneX database to test and refine performance
- Develop APIs to the database for ease in testing new analysis methods

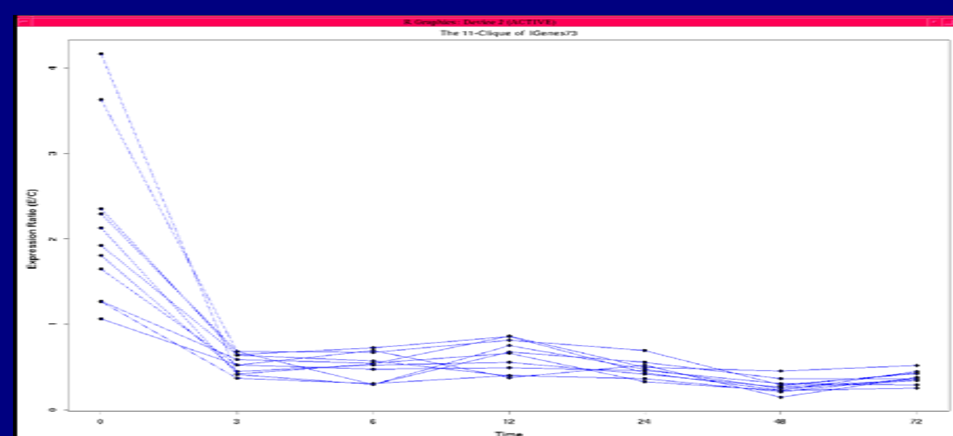
Collaborators

- CalTech, UVa, VCU

Sponsors (Current)

- NSF





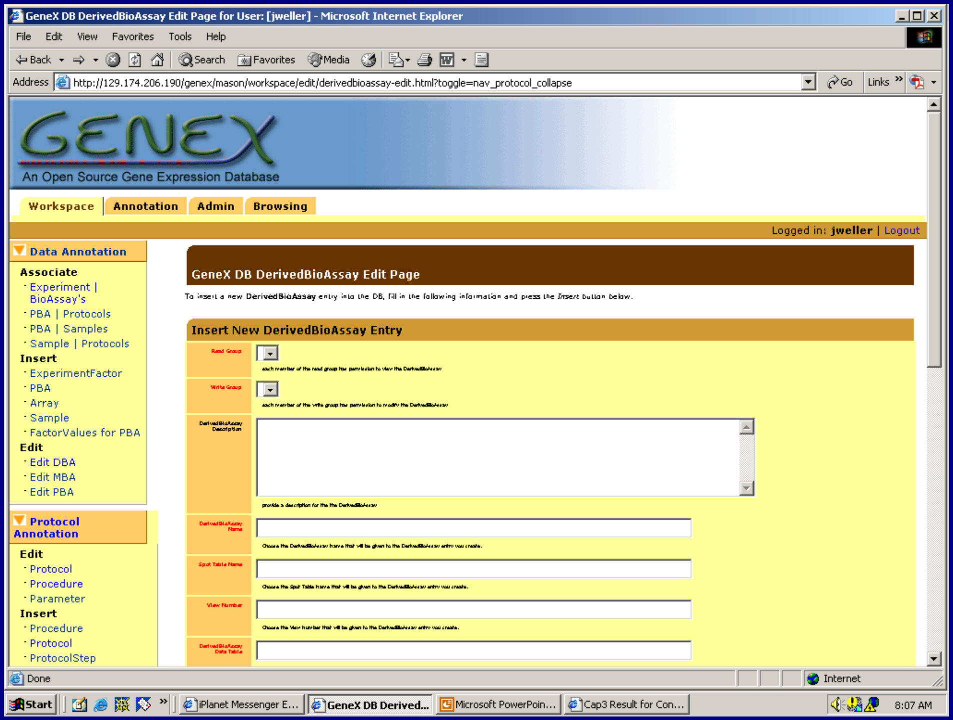
http://genex.sourceforge.net/genexml.shtml

Workspace Concept

GeneX has Project- and Investigator-specific Views, that we call Workspaces.

It is possible to define new Tables for the Workspace, allowing greater flexibility for storing the results of computational experiments.

We are working on providing code that will generate basic Reports from these new tables automatically.



Academic Programs at GMU
www.gmu.edu

- Bioinformatics and Computational Biology
  - Ph.D. in Bioinformatics (1992)
  - M.S. in Bioinformatics (2002)
- www.binf.gmu.edu

- Computational Science and Informatics
  - Ph.D. in Computational Neurobiology
  - M.S. in CSI
  - Certificate in CSI

- Environmental Sciences and BioSciences (CAS)
  - Ph.D in Environmental Sciences and Policy
  - Ph.D. in BioSciences
  - M.S. in Biology
- mason.gmu.edu/~esp